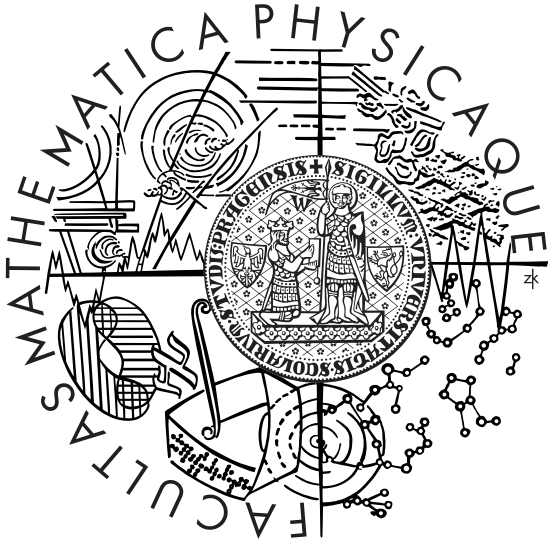
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

# BAKALÁŘSKÁ PRÁCE




Marcel Beno

## P-splajnová regrese

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce : Mgr. Jozef Juríček, M.Sc.

Studijní program : Matematika , Finanční matematika

2010

Na tomto mieste by som rád vyjadril vďaku svojmu vedúcemu práce Mgr. Jozefovi Juríčkovi, M.Sc. za priateľský prístup a za zaujímavú tému, ktorú som mohol vďaka nemu spracovať.

Taktiež by som chcel poďakovať Univerzite Karlovej v Praze za možnosť štúdia na tejto významnej škole, za to, že rozšírila moje matematické vnímanie, moju náklonnosť k matematike a predovšetkým za tie nádherne roky, ktoré som mohol vďaka štúdiu stráviť v Prahe.

Z osôb na tejto škole patrí najväčšia vďaka mojej študijnej referentke, pani Marcele Všechovskej, za jej trpezlivosť a ochotu počas môjho štúdia.

Nakoniec by som rád poďakoval ešte svojej strednej škole Gymnáziu Grősslingová v Bratislave za vynikajúci prístup k študentom a podporu matematiky, vďaka ktorej som sa rozhodol pre štúdium na matematicko-fyzikálnej fakulte v Prahe.

Prehlasujem, že som svoju bakalársku prácu napísal samostatne a výhradne s použitím citovaných prameňov. Súhlasím so zapožičiavaním práce a jej zverejňovaním.

V Prahe dňa 5.8.2010 Marcel Beno

# Obsah

Názov práce : P-splajnová regrese
Autor : Marcel Beno
Katedra : Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce : Mgr. Jozef Juríček, M.Sc.
e-mail vedoucího : jozef.juricek@matfyz.cz

**Abstrakt :** V predloženej práci bude študovaná teória P-splajnovej regresie, jej prednosti oproti klasickej B-splajnovej regresii a možnosti jej využtia v praxi. Teoretická časť sa opiera hlavne o práce Eilersa a Marxa (1996) v oblasti P-splajnovej regresie, Bollaertsa s kolektívom (2006) a Zhanga (2004). Následne je teória implementovaná do softvérového programu a otestovaná v simulačnej štúdii s umelo vytvorenými dátami. V praktickej časti sú získané poznatky použité v reálnej štúdii výskytu Parvovirusu B19 u populácie, kde sú skúmané výsledky sérologických testov ako funkcie veku pacientov v dobe vykonávania daného testu.

Kľúčové slová : Regresia , B-splajn , P-splajn , vyhladzovanie dát

Title : P-spline regression
Author : Marcel Beno
Department : Department of probability and mathematical statistics
Supervisor : Mgr. Jozef Juríček, M.Sc.
Supervisor's e-mail adress : jozef.juricek@matfyz.cz

**Abstract :** In the work we study the theory of P-spline regression, its advantages compared to common B-spline regression and its practical use. Theoretical part of the work is based mainly on works of Eilers and Marx (1996) in the field of the P-spline refression, Ballaers and company (2006) and Zhang (2004). In the second part we implement a computer program. The software provides us with a simulation study of the theory based on an artificial data. In the final practical part we apply the theory and study the presence of the parvovirus B19. The result of a serological test, the appearance of the virus, is considered to be a function of the age of a patient in the time of testing.

Keywords : Regression , B-spline , P-spline , smoothing

# Kapitola 1

# Úvod

V súčasnej dobe sa stáva stále vo väčšej miere dôležitá grafická prezentácia dát, najmä kvôli svojej jednoduchosti a prehľadnosti. Na tento účel sa využívali parametrické modely, tie však prinášali so sebou problém voľby "správnych" parametrov. Navyše, množstvo súborov dát je príliš obsiahlych a komplexných na to, aby mohli byť vhodne popísané parametrickými modelmi a aj preto je vhodnejšie používať modely neparametrické. Ako vhodné sa ukazuje modelovanie pomocou takzvaných B-splajnov, ktorých základy a metodológia budú vysvetlené v sekcii 2.1 .

Neparametrický prístup, ktorý spočíva v rozšírení B-splajnov o penalizáciu za hladkosť, navrhli vo svojej práci Eilers a Marx (1996). Tento prístup sa nazýva P-splajnová regresia a jeho princíp aj s ďalšími doplnkami sa uvedie v sekcii 2.2.

Celá 2. kapitola, ktorá bude venovaná teórii, bude doplnená o grafické ilustrácie.

V 3. kapitole sa metódy aplikujú na umelo vytvorené dáta a bude sa skúmať, či ich teoretické vlastnosti zodpovedajú realite.

Aplikácia metód na reálne dáta bude predvedená v 4. kapitole, spolu s popisom dát a konkrétnymi závermi. Ako dáta sú použité výsledky testov Parvovírusu B19, ktorý bude podrobnejšie popísaný v sekcii 4.1. V sekcii 4.2 budú dáta popísané niektorými základnými štatistikami a následne, v sekcii 4.3, na ne bude aplikovaná P-splajnová regresia.

Aby mohli byť tieto metódy použité v praxi, boli implementované do štatistického softvéru. Pre tento účel bol použitý program Mathematica 7, ktorý je jedným z najviac používaných v štatistických kruhoch pre svoju komplexnosť. V 5. kapitole budú popísané niektoré problémy, ktoré môžu pri tomto procese nastať a aké základné funkcie sa využívali pri implementácii.

Zdrojové kódy funkcií z programu Mathematica 7 sú pre užívateľov dostupné v prílohe A.

## 1.1 Model

Pre jednoduchosť sa práca obmedzuje na modely s jednou nezávislou premennou a vo všeobecnosti sa predpokladá, že :

$$y_i = f(x_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathbf{N}(0, \sigma^2), \quad i = 1, \dots N, \tag{1.1}$$

kde $x_i$ je nezávislá premenná (regresor), $y_i$ je závislá premenná, $\varepsilon_i$ vyjadruje "chyby v pozorovaní", $N$ značí počet napozorovaných dát a $f$ regresnú funkciu.

Ďalej je zadaný predpoklad hladkosti regresnej funkcie :

$$f \in C^q(D), \quad q \in \mathbb{N},$$

$f$ je $q$-krát spojito derivovateľná na definičnom obore $D = [x_l, x_r]$; $x_l$ je najmenšia možná hodnota $x$ , $x_r$ najvyššia.

# Kapitola 2

# P-splajnová regresia

Táto kapitola je venovaná teórii a bude podrobne vysvetlená metodológia neparametrických P-splajnov. V skutočnosti je pomenovanie týchto modelov ako neparametrických nie úplne presné, vzhľadom na to, že jednotlivé splajny sú popísané parametrami, aj keď je ich mnoho. Vhodnejším názvom pre tieto modely by bolo napríklad mnohoparametrické, alebo semiparametrické. V sekcii 2.1 bude poskytnutý náhľad do parametrického modelovania v podobe B-splajnovej regresie. Táto regresia využíva metódu najmenších štvorcov, ktorá bude v sekcii tiež prednesená. Samotná neparametrická P-splajnová regresia, ktorá je v podstate regresiou pomocou metódy najmenších štvorcov s veľkým počtom B-splajnov a jej rozšírením o penalizáciu za *malú hladkosť*, sa bude nachádzať v sekcii 2.2. V druhej časti sekcie budú následne popísané možné spôsoby výpočtu tejto penalizácie a model bude skompletizovaný určením intervalu spoľahlivosti regresného odhadu.

Rovnako ako v časti 1.1 sa predpokladá model:

$$y_i = f(x_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathbf{N}(0, \sigma^2), \quad i = 1, \dots N, \tag{2.1}$$

kde platí: $f \in C^{q-1}(D)$ a $D = [x_l, x_r]$.

## 2.1 B-splajny

Základné informácie a definície nutné pochopeniu metódy B-splajnov budú popísané v úvode tejto časti.

**Definícia 2.1** (B-spline). B-splajnom rádu $q$ ($q \in \mathbb{N}$) s *vnútornými uzlami* $\lambda_1 < \lambda_2 < \ldots < \lambda_q$ a hranicami (resp. vonkajšími uzlami) $\lambda_0, \lambda_{q+1}$, takými, že $\lambda_0 < \lambda_1$a $\lambda_q < \lambda_{q+1}$, rozumieme po častiach ( na intervaloch $(\lambda_i, \lambda_{i+1}), i = 0, \ldots, q$) nezápornú polynomiálnu funkciu $B(.)$ stupňa $q$, ktorá je konštantne nulová na množine$(\lambda_0, \lambda_{q+1})^C$a pre ktorú platí $B \in C^{q-1}(\mathbb{R})$.

*Poznámka 2.2.* $\lambda_0$ bude odteraz pomenovávaný ako *ľavý uzol*.

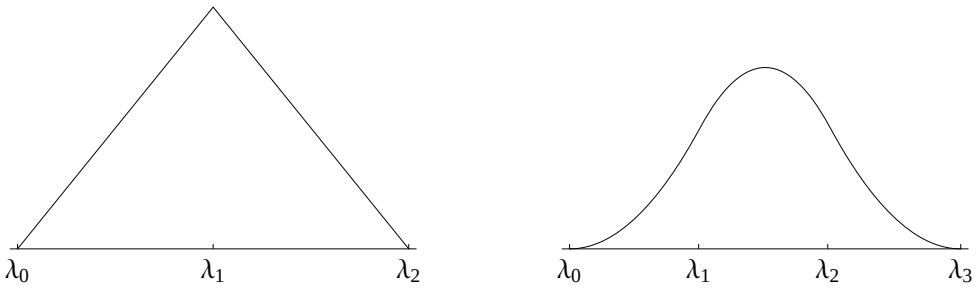
B-splajn rádu $q$ je teda polynomiálna funkcia (polynómy sú stupňa $q$) taká, že:

$B(\lambda_{i+}) = B(\lambda_{i-}); \quad B'(\lambda_{i+}) = B'(\lambda_{i-}); \ldots; \quad B^{(q-1)}(\lambda_{i+}) = B^{(q-1)}(\lambda_{i-}); i = 1, ..., q.$

A pre krajné body platí:

$B(\lambda_0) = B(\lambda_{q+1}) = B'(\lambda_0) = B'(\lambda_{q+1}) = \ldots = B^{(q-1)}(\lambda_0) = B^{(q-1)}(\lambda_{q+1}) = 0.$

Jednoduchý príklad je uvedený na obrázku 1, kde sú ukázané jednotlivé B-splajny rádu $q = 1$ a $q = 2$ .




B-splajn rádu 1 B-splajn rádu 2

Obr. 1 : Príklady B-splajnov

Splajn rádu 1 teda pozostáva z dvoch lineárnych častí. Prvá z $\lambda_0$ do $\lambda_1$ a druhá z $\lambda_1$do $\lambda_2$. Uzly sú $\lambda_0$, $\lambda_1$ a $\lambda_2$ . Na intervale $(\lambda_0, \lambda_2)^C$ je potom tento B-splajn nulový. V druhej časti obrázku je ukázaný splajn rádu 2, ktorý pozostáva z troch kvadratických častí spojených v dvoch uzloch. V spájaných uzloch sa rovnajú nielen funkčné hodnoty, ale aj hodnoty prvých derivácií týchto polynómov. B-splajn rádu 2 je následne tvorený 4 uzlami $\lambda_1$, $\lambda_2$, $\lambda_3$ a $\lambda_4$.

**Definícia 2.3.** (B-splajnová báza) Pre daný interval $[x_l, x_r] \subset \mathbb{R}$ a $n, q \subset \mathbb{N}$ , pričom $n > q$ , značíme :

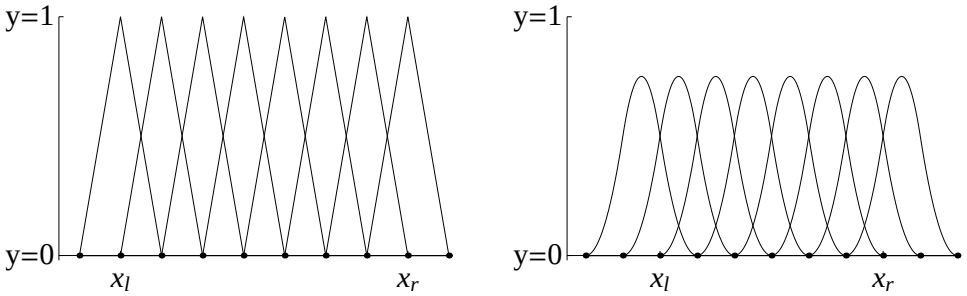
$$n' := n - q$$
$$h := \frac{x_r - x_l}{n'}$$
$$t_1 := x_l - qh$$
$$t_k := t_1 + (k-1)h$$

Potom *B-splajnová báza* na intervale $[x_l; x_r]$, s dimenziou $n$, ekvidištantnými uzlami a rádom $q$ je sada B-splajnov takých, že platí:

$$\forall x \in [x_l; x_r] : \sum_{j=1}^{n} B_j(x, q) = 1 \tag{2.2}$$

Kde $B_j(x, q)$ značí hodnotu B-splajn rádu $q$ s ľavým uzlom $t_j$ (j-ty splajn) v bode $x$. V prípade, že $q$=0, je hodnota $B_j(x, 0)$ rovná 1 pre každý bod v intervale $[x_l; x_r]$ a nulová mimo tohoto intervalu.

Týmto spôsobom je možné pre väčšie množstvo uzlov skonštruovať ľubovolné množstvo B-splajnov (viz obrázok 2 pre splajny rádu 1 a 2 ).




B-splajnová báza rádu 1 B-splajnová báza rádu 2

Obr. 2 : Príklady B-splajnových báz

Ako vidno B-splajny sa prekrývajú. Splajny rádu 1 sa prekrývajú s dvoma splajnami, druhého rádu so štyrmi a tak ďalej ( to samozrejme neplatí o krajných splajnoch, ktoré sa prekrývaju s menším počtom). B-splajny na obrázku 2 sú B-splajnami s ekvidištantnými (rovnako vzdialenými) uzlami.

V skutočnosti môžu existovať aj splajny s nerovnako vzdialenými uzlami, v tejto práci (rovnako tak aj pri analýze dát) nebudú ale brané v úvahu.

De Boor (1978) ponúka algoritmus na výpočet derivácií B-splajnov a taktiež spôsob ich konštrukcie pomocou splajnov nižších rádov . Vďaka faktu, že splajn rádu nula je konštantná funkcia na intervale $[x_l; x_r]$, je veľmi jednoduché počítať splajny akéhokoľvek rádu:

$$B_j(x,q)=\frac{x-t_j}{qh}B_j(x,q-1)+\frac{x-t_{j+q-1}}{qh}B_{j+1}(x,q-1) \tag{2.3}$$

$$B_{j+k}(x,q)=B_j(x-kh,q) \tag{2.4}$$

$$h\sum_j a_j B_j^{'}(x,q)=\sum_j a_j B_j(x,q-1)-\sum_j a_{j+1}B_{j+1}(x,q-1)=-\sum_j \Delta a_{j+1}B_j(x,q-1), \tag{2.5}$$

pričom $\Delta a_j$ je diferencia prvého rádu: $\Delta a_j = a_j - a_{j-1}$. Indukciou potom:

$$h^2\sum_j a_j B_j^{''}(x,q)=\sum_j \Delta^2 a_j B_j(x,q-2), \tag{2.6}$$

kde $\Delta^2 a_j = \Delta\Delta a_j = a_j - 2a_{j-1} + a_{j-2}$ je diferencia druhého rádu. Pre vyššie rády analogicky platí: $\Delta^k a_j = \underbrace{\Delta...\Delta}_{k} a_j$.

*Poznámka 2.4.* V ďalšom texte bude používaný skrátený zápis B-splajnov $B_j(x) \equiv B_j(x,q)$.

Odhad regresnej funkcie $f$ bude následne konštruovaný pomocou metódy najmenších štvorcov.

**Metóda 2.5.** (B-splajnová Regresia Metódou Najmenších Štvorcov). Odhad regresnej funkcie v modeli (1.1) pomocou metódy najmenších štvorcov je počítaný ako lineárna kombinácia B-splajnov z B-splajnovej bázy:

$$\hat{f}(x_i) \equiv f(x_i,\hat{a}) \equiv \hat{a}_1 B_1(x_i) + \hat{a}_2 B_2(x_i) + \ldots + \hat{a}_n B_n(x_i)\ ,\ i = 1,\ldots,N\ ,$$

čo si možno zapísať v maticovom tvare ako:

$$\begin{pmatrix} \hat{f}(x_1) \\ \vdots \\ \hat{f}(x_N) \end{pmatrix} = \begin{pmatrix} B_1(x_1) & \ldots & B_n(x_1) \\ \vdots & \ldots & \vdots \\ B_1(x_N) & \ldots & B_n(x_N) \end{pmatrix} \begin{pmatrix} \hat{a}_1 \\ \vdots \\ \hat{a}_n \end{pmatrix}$$

$$\hat{f}(x) = \boldsymbol{B}\hat{a},$$

kde: $\boldsymbol{B} = (b_{ij})_{i=1,\ldots,N;j=1,\ldots,n}$ a $(b_{ij}) = B_j(x_i)$.

Následne sa minimalizuje takzvaná stratová funkcia $S$ ( kvadratická odchýlka voči dátam) :

$\hat{a} \equiv (\hat{a}_1 \dots \hat{a}_n)^{\mathrm{T}} = \arg\min_{a\in\mathbb{R}^n} S(a)$, $\qquad S(a)=\sum_{i=1}^{N} (y_i - f(x_i,a))^2$,
kde $f(x,a) = \sum_{j=1}^{n} a_j B_j(x)$.

$S(a)$ je možné si ďalej upraviť na : $S(a)=(y-\boldsymbol{B}a)^{\mathrm{T}}(y-\boldsymbol{B}a)$ , kde $y=(y_1,\dots,y_N)^{\mathrm{T}}$ a následne:

$$S(a) = (y-\boldsymbol{B}a)^T(y-\boldsymbol{B}a) = (y-\boldsymbol{B}\hat{a}+\boldsymbol{B}\hat{a}-\boldsymbol{B}a)^T(y-\boldsymbol{B}\hat{a}+\boldsymbol{B}\hat{a}-\boldsymbol{B}a)= \\ = S(\hat{a}) + (y-\boldsymbol{B}\hat{a})^T\boldsymbol{B}(\hat{a}-a) + (\hat{a}-a)^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}(y-\boldsymbol{B}\hat{a}) + (\hat{a}-a)^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}(\hat{a}-a).$$

V prípade, že:

$$\boldsymbol{B}^{\mathrm{T}}(y-\boldsymbol{B}\hat{a})=\boldsymbol{B}^{\mathrm{T}}y - \boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}\hat{a} = 0, \tag{2.7}$$

platí:

$$S(a) = S(\hat{a}) + \|\boldsymbol{B}(\hat{a}-a)\|^2 \tag{2.8}$$

Ak teda $\hat{a}$ je riešením (2.7), je aj bodom minima funkcie S. $\tilde{a}$ buď bodom globálneho minima funkcie S. Potom z rovnice (2.8), vyplýva:

$$\min S = S(\tilde{a}) = S(\hat{a}) + \|B(\hat{a}-\tilde{a})\|^2 = S(\hat{a}) \Rightarrow B\tilde{a} = B\hat{a}.$$

Takže každý bod globálneho minima funkcie S splňuje rovnicu (2.7) a vyhladené data $B\tilde{a}$ nezávisia na volbe $\hat{a}$.

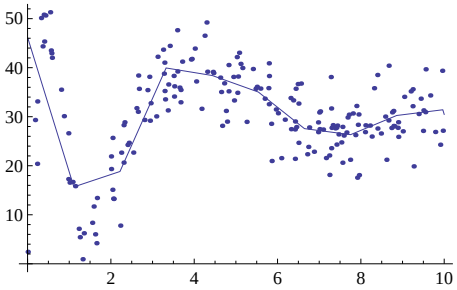
Hodnosť rozšírenej matice sústavy (2.7) je podľa vety o dimenzii jadra a obrazu rovná:

$$h(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}, \boldsymbol{B}^{\mathrm{T}}y) = n - \dim\left\{a\in\mathbb{R} : a^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}=0, a^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}y=0\right\} \\ n - \dim\left\{a\in\mathbb{R}^n : \boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}a=0, \boldsymbol{B}a=0\right\} \\ n - \dim\left\{a\in\mathbb{R}^n : \boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}a=0\right\} = h(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}),$$

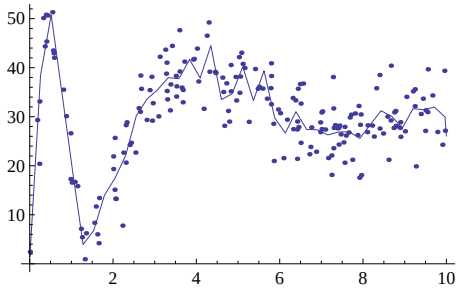
pretože hodnosť $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}$ je rovnaká ako hodnosť matice $\boldsymbol{B}$ a teda $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}a=0$ práve keď $\boldsymbol{B}a=0$. Existuje teda aspoň jedno riešenie $\hat{a}$ sústavy (2.7).

Ak je matica $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}$ regulárna, potom vyrovnané dáta môžme získať zo vzťahu $\boldsymbol{B}\hat{a} = \boldsymbol{H}y$, kde $\boldsymbol{H}$ je projekčná matica: $\boldsymbol{H} = \boldsymbol{B}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}$.

Na obrázku 3a a 3b sú uvedené príklady B-splajnov pre rôzne hodnoty hodnoty $n$ a $q$.

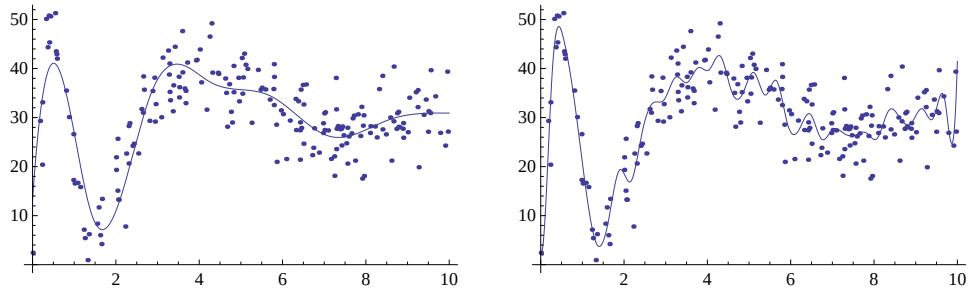



B-splajnová regresia($q=1, n=10$)




B-splajnová regresia($q=1, n=40$)

Obr. 3a: Príklady B-splajnovej regresie




B-splajnová regresia($q = 3, n = 10$) B-splajnová regresia($q = 3, n = 40$)

Obr. 3b: Príklady B-splajnovej regresie

*Poznámka 2.6.* Dáta na všetkých grafoch sú rovnaké a sú umelo vytvorené. Presný popis ich konštrukcie bude uvedený v kapitole 3.

Pre daný počet splajnov $n$ a ich rád $q$ nie je teda výpočetne náročné spočítať odhad regresnej funkcie pomocou metódy najmenších štvorcov. Hlavným problémom B-splajnov ostáva teda výber týchto dvoch hodnôt. V praxi sa postupne začali využívať najmä splajny tretieho rádu, ktoré pri postačujúcom počte splajnov poskytujú dostatočnú flexibilitu. U počtu B-splajnov vedie malá hodnota $n$ k tomu, že regresná funkcia je moc hladká a dochádza k strate dôležitých informacii. Na druhej strane veľký počet B-splajnov spôsobí, že krivka je veľmi flexibilná a zahŕňa v sebe aj nahodné fluktuácie, čo je nežiaduci efekt.

## 2.2 P-splajny

Jedno z možných riešení tohoto problému uviedol vo svojej práci O'Sullivan (1986). Navrhol použitie veľkého množstva B-splajnov spolu so *zjemňujúcou* penalizáciou $P$ (resp. s penalizáciou za prílišnú hladkosť), pozostávajúcou z integrácie štvorca druhej derivácie vyhladenej krivky. Táto penalizácia má za cieľ korigovať veľkú flexibilitu týchto splajnov a stala sa štandardom v literatúre zaoberajúcej sa problematikou splajnov, aj keď je možné použiť derivácie iných rádov. Použitie derivácie prvého rádu vedie k jednoduchým rovniciam a po častiach lineárnemu vyhladeniu, pričom použitie vyšších derivácii vyústi v zložitý systém rovníc a veľmi jemné vyhladenie.

$$P = P(\lambda) = \lambda \int_{x_l}^{x_r} \left[ (f^{(k)}(x, a) \right]^2 \tag{2.9}$$

**Metóda 2.8.** (O'Sullivan: P-splajnová Regresia). Nech $k < q$. Odhad regresnej funkcie $f$ v modeli (1.1) (so zadanými hodnotami vyhladzovacej konštanty $\lambda \geq 0$ a $k$-tym stupňom penalizácie) je lineárna kombinácia B-splajnov z B-splajnovej bázy $\hat{f}_{k,\lambda}(x) \equiv f(x, \hat{a}_{k,\lambda}) \equiv \sum_{j=1}^{n} \hat{a}_{j_{k,\lambda}} B_j(x)$ pričom:

$$\hat{a_{k,\lambda}} = \arg\min_{a \in \mathbb{R}^n} S_{k,\lambda}(a), \qquad S_{k,\lambda}(a) = \sum_{i=1}^{N} (y_i - f(x_i, a))^2 + \lambda \int_{x_l}^{x_r} \left[ (f^{(k)}(x, a) \right]^2, \tag{2.10}$$

kde $f(x, a) = \sum_{j=1}^{n} a_j B_j(x)$.

Prístup O'Sullivana (1986) bol inšpirovaný metodológiou *vyhladzovacích splajnov.* Viac o tejto metóde je možné nájsť v Green (1994).

Eilers a Marx (1996) navrhli taktiež použitie veľkého množstva B-splajnov s ekvidištantými uzlami, ale s diskrétnou penalizáciou $P_\Delta$ založenou na diferenciách koeficientov B-splajnov z bázy:

$$P_\Delta{=}P_\Delta(\lambda) = \lambda \sum_{j=k+1}^{n} \left(\Delta^k a_j\right)^2 \tag{2.11}$$

**Metóda 2.9.** (Eilers: P-splajnová Regresia). Nech $k < q$. Odhad regresnej funkcie $f$ v modeli (1.1) (so zadanými hodnotami vyhladzovacej konštanty $\lambda \geq 0$ a $k$-tym stupňom penalizácie) je lineárna kombinácia B-splajnov z B-splajnovej bázy $\hat{f}_{k,\lambda}(x) \equiv f(x, \hat{a}_{k,\lambda}) \equiv \sum_{j=1}^{n} \hat{a}_{j_{k,\lambda}} B_j(x)$ pričom:

$$\hat{a}_{k,\lambda} = \arg\min_{a\in\mathbb{R}^n} S_{k,\lambda}(a), \qquad S_{k,\lambda}(a){=}\sum_{i=1}^{N}(y_i - f(x_i, a))^2 {+} \lambda \sum_{j=k+1}^{n} \left(\Delta^k a_j\right)^2. \tag{2.12}$$

Teda rovnako ako v prípade B-splajnovej regresie sa hľadá minimum funkcie $S_{k,\lambda}(a)$. V tejto kapitole bude ponúknutý spôsob získania optimálnej hodnoty $\hat{a}_{k,\lambda}$ pomocou derivácie $S_{k,\lambda}(a)$.

Vzťah (2.12) si možno prepísať:

$$S(a){=}(y - \boldsymbol{B}a)^{\mathrm{T}}(y - \boldsymbol{B}a) + \lambda \boldsymbol{D}_k^{\mathrm{T}} \boldsymbol{D}_k a, \tag{2.13}$$

kde $\boldsymbol{D}_k$ je matica reprezentujúca diferencie $\triangle^k$.

Minimum funkcie $S(a)$ sa nachádza v bode $\hat{a}$, práve keď :

$$\nabla S(\hat{a}){=}\left(\frac{\partial S(\hat{a})}{\partial \hat{a}_1}, ..., \frac{\partial S(\hat{a})}{\partial \hat{a}_n}\right)^{\mathrm{T}} {=}0 \tag{2.14}$$

Z (2.13) a (2.14) vyplýva:

$$\boldsymbol{B}^{\mathrm{T}} y{=}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}\hat{a} + \lambda \boldsymbol{D}_k^{\mathrm{T}} \boldsymbol{D}_k \hat{a} = (\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} + \lambda \boldsymbol{D}_k^{\mathrm{T}} \boldsymbol{D}_k)\hat{a}$$

V prípade, že hodnosť matice $(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} + \lambda \boldsymbol{D}_k^{\mathrm{T}} \boldsymbol{D}_k) = n$ sa $\hat{a}$ vyjadrí ako:

$$\hat{a} = (\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} + \lambda \boldsymbol{D}_k^{\mathrm{T}} \boldsymbol{D}_k)^{-1} \boldsymbol{B}^{\mathrm{T}} y$$

Tento prístup je veľmi podobný tomu, ktorý ponúkol O'Sullivan, pretože užitím derivácií B-splajnov podľa (2.6) platí (Eilers a Marx 1996):

$$h^2 P_2 = \lambda \int_{x_l}^{x_r} \left\{ \sum_j a_j B_j^{''}(x, 3) \right\}^2 dx = \lambda \int_{x_l}^{x_r} \left\{ \sum_j a_j B_j^{''}(x, 1) \right\}^2 dx,$$

čo je možné si prepísať ako:

$$h^2 P_2 = \lambda \int_{x_l}^{x_r} \left\{ \sum_j \sum_k \Delta^2 a_j \Delta^2 a_k B_j(x, 1) B_k(x, 1) \right\}^2 dx.$$

B-splajny rádu 1 sa prekrývajú len ak $j = k - 1$, alebo $j{=}k + 1$, a teda väčšina násobkov $B_j(x,1)B_k(x,1)$ je nulových. Vďaka tomu je možné si vzťah ďalej upravovať až do výsledného stavu:

$$h^2 P_2 = \lambda \int_{x_l}^{x_r} \left[ \left\{ \sum_j a_j B_j(x,1) \right\}^2 + 2 \sum_j \Delta^2 a_j \Delta^2 a_{j-1} B_j(x,1) B_{j-1}(x,1) \right] dx$$

$$h^2 P_2 = \lambda \sum_j \left(\Delta^2 a_j\right)^2 \int_{x_l}^{x_r} \sum_j B_j^2(x,1) dx + 2\lambda \sum_j \Delta^2 a_j \Delta^2 a_{j-1} \int_{x_l}^{x_r} B_j(x,1) B_{j-1}(x,1) dx.$$

V zjednodušenom zápise potom:

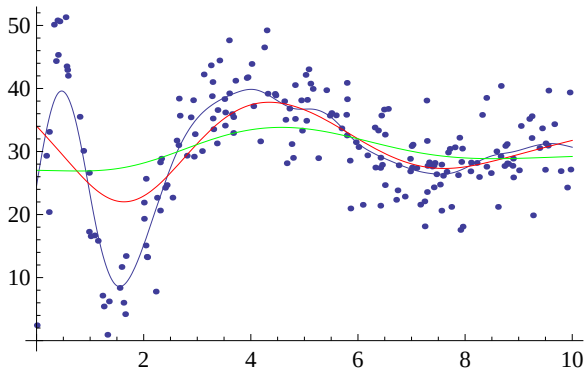
$$h^2 P_2 = \lambda \left\{ c_1 \sum_j \left(\Delta^2 a_j\right)^2 + c_2 \sum_j \Delta^2 a_j \Delta^2 a_{j-1} \right\}, \tag{2.15}$$

kde $c_1$a $c_2$ sú konštanty pre dané (ekvidištantné) uzly:

$$c_1 = \int_{x_l}^{x_r} B_j^2(x,1) dx, c_2 = \int_{x_l}^{x_r} B_j(x,1) B_{j-1}(x,1) dx,$$

Prvá časť vzorca (2.15) zodpovedá vyhladzovacej penalizácii $P_\Delta$ druhého rádu. Druhá časť vedie k omnoho komplexnejším rovniciam pri výpočte minima $S(a)$ v (2.10). Táto komplexnosť vyplýva z prekrývania sa B-splajnov. Pri použití v praxi s B-splajnami a diferenciami vyšších rádov stúpa rapídne výpočetná náročnosť a je zložité naprogramovať túto penalizáciu do automatickej procedúry na výpočet regresnej funkcie $f$. S použtím penalizácie $P_\Delta$ tento problém odpadá.

Na obrázku 4 sú znázornené P-splajnové regresie pre rôzne parametre $\lambda$.




$(q = 3, n = 30, k = 2, \lambda = 1, \lambda = 10, \lambda = 100)$

Obr. 4: P-splajnová regresia pre rôzne vyhladzovacie parametre

Zavedením P-splajnov bol problém určenia optimálnej dimenzie $n$ (výber z prirodzených čísel - *diskrétnej množiny*) "prenesený" na problém určenia optimálnej hodnoty vyhladzovacej konštanty $\lambda \geqq 0$ (z kladných reálnych čísel - *kontinua*)

Na parameter $\lambda$ je možné nahliadať aj ako na tzv. *vyhladzovací parameter*. Problém jeho výberu je jedným zo základných problémov v neparametrickej štatistike. Často používanými metódami na získanie optimálnej hodnoty $\lambda$ sú *Akaikeho informačné kritérium* (AIC) a metóda *krížového overovania*. V ďalšom texte bude podrobnejšie popísaná a následne aj použitá na dátach druhá z týchto dvoch metód.

*Poznámka 2.10.* Pojem *krížové overovanie* nie je v slovenčine (resp. v češtine) zaužívaný a v ďalšom texte bude používaný anglický názov metódy *cross-validation*.

**Metóda 2.11.** (Cross-Validation, Generalized Cross-Validation). Optimálny odhad vyhladzovacej konštanty metódou cross-validation sa spočíta zo vzťahu:

$$\hat{\lambda}_{CV} = \arg\min_{\lambda \geqq 0} CV(\lambda), \quad CV(\lambda) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{f}_\lambda^{(-i)}(x_i)\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\frac{\left(y_i - \hat{f}_\lambda(x_i)\right)^2}{(1-h_{ii})^2}, \tag{2.16}$$

kde $h_{ii}$ je $i$-ty diagonálny prvok projekčnej matice $\boldsymbol{H}=(h_{im})_{i,m=1,\ldots,N}$:

$$\boldsymbol{H} = \boldsymbol{B}\left(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} + \lambda \boldsymbol{D}_k^{\mathrm{T}}\boldsymbol{D}_k\right)^{-1}\boldsymbol{B}^{\mathrm{T}}$$

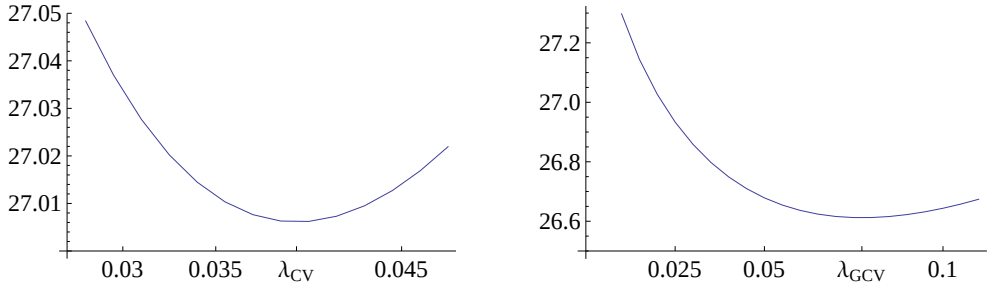
a $\hat{f}_\lambda^{(-i)}$ je odhad funkcie $f$ pomocou P-splajnovej regresie (s daným $\lambda$) pre dáta $(x_j, y_j)_{j=1\ldots,i-1,i+1,\ldots,N}$ ($i$-te pozorovanie je vynechané)

Wahba(1990) ponúkol taktiež model tzv. *Generalized cross-validation* :

$$\hat{\lambda}_{GCV} = \arg\min_{\lambda \geq 0} GCV(\lambda), \quad GCV(\lambda) = \sum_{i=1}^{N}\frac{\frac{1}{N}\left(y_i - \hat{f}_\lambda(x_i)\right)^2}{(1-\frac{tr(H)}{N})^2}, \tag{2.17}$$

*Poznámka 2.12.* značí sa: $\hat{y} = \left(\hat{f}(x_i)\right)_{i=1,\ldots,N} = \boldsymbol{B}\hat{a} = \boldsymbol{H}y$.

Graf CV a GCV funkcií je zobrazený na obrázku 5. Pre budúce úlohy boli vyčíslené $\lambda_{CV} = 0.0393$ a $\lambda_{GCV} = 0.0773$ pre tieto dáta.



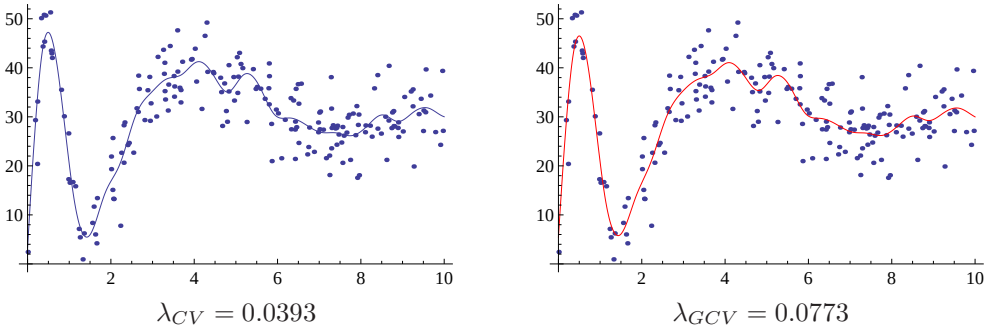

$CV(\lambda)$ $GCV(\lambda)$

Obr. 5: Graf CV a GCV funkcií

Rozdiely u oboch metód sú vo všeobecnosti veľmi malé. Ako je uvedené v definícii, najvhodnejší odhad $\lambda$ sa získa minimalizovaním $CV(\lambda)$ resp. $GCV(\lambda)$ a následne sa pomocou neho zostrojí *optimálny P-splajnový odhad* regresnej funkcie $f$.

**Metóda 2.12.** Pre o*ptimálny P-splajnový odhad* regresnej funkcie $f$ v modeli (1.1) platí: $\hat{f} = \hat{f}_{\hat{\lambda}}$,kde $\hat{\lambda} = \hat{\lambda}_{CV}$, alebo $\hat{\lambda} = \hat{\lambda}_{GCV}$.

Na obrázku 6 bola aplikovaná na dáta P-splajnová regresia s vyhladzovacími konštantami spočítanými metódami CV a GCV




$(q = 3, n = 30, k = 2)$

Obr. 6: P-splajnová regresia s metódou CV a GCV

Na skompletovanie metódy a určenie intervalu spoľahlivosti regresnej funkcie $f$ je potrebné ešte konzistentne odhadnúť rozptyl $\sigma^2$ v modeli (1.1). Rozdiely reziduí pri optimálnej hodnote $\lambda$ sú vhodnou voľbou pre tento výpočet (Juríček 2009):

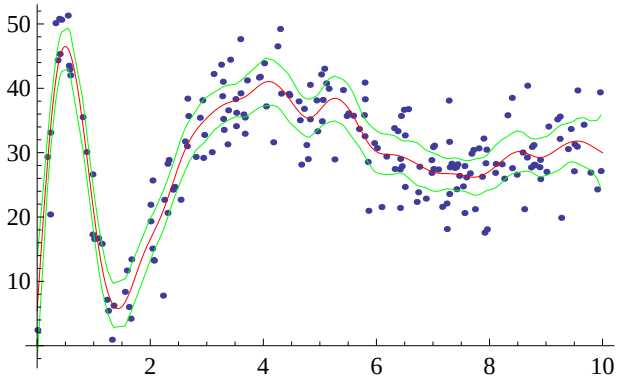
$$\sigma^2 = \frac{RSS}{N - tr(H)} = \frac{\sum_{i=1}^{N} \hat{e}_i^2}{N - tr(H)} = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N - tr(H)} \tag{2.18}$$

potom, (1-$\alpha$) $\times$100 % interval spoľahlivosti (Wahba 1990):

$$\hat{f}(x_i) \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sigma \sqrt{h_{ii}}, i = 1, ..., N \tag{2.19}$$

$\Phi^{-1}(.)$ je kvantilová funkcia normovaného normálneho rozdelenia N(0,1).

Na obrázku 7 je interval spoľahlivosti pre $\alpha = 0.05$ v metóde GCV




$(q = 3, n = 30, k = 2, \lambda_{GCV} = 0.0773, \alpha$=0.05)

Obr. 7: Interval spoľahlivosti pre P-splajnovú regresiu

# Kapitola 3

# Simulačná štúdia

Za účelom skúmania ako sa navrhnuté metódy "správajú", bola v tejto kapitole vykonaná simulačná štúdia. Umelo vytvorené dáta boli náhodne generované z modelu (1.1). Funkcia $f_D(x)$, pomocou ktorej boli dáta generované, bola zvolená tak, aby zodpovedala funkciám s ktorými sa možno stretnúť v praxi - napríklad lekárstve. Skonštruovaných bolo viacero sérií dát, na ktorých boli po aplikácii P-splajnovej regresie ($n$=30, $q = 3, k = 2$) skúmané výberové stredné hodnoty a rozptyly pre vyhladzovacie konštanty a rozptyly regresných funkcii. Pri voľbe $\lambda$sa využívala metóda generalized cross-validation.

Jedna séria dát obsahovala dáta o rozsahu 500 prvkov a $f_D(x)$ mala definičný obor $[0, 10]$. Týchto sérií bolo skonštruovaných $u = 100$ a následne sa pre každú z týchto sérií pozorovali hodnoty ${}_D\lambda_{j,cv}$ a ${}_D\sigma_j^2$, $j = 1, ..., 100$.

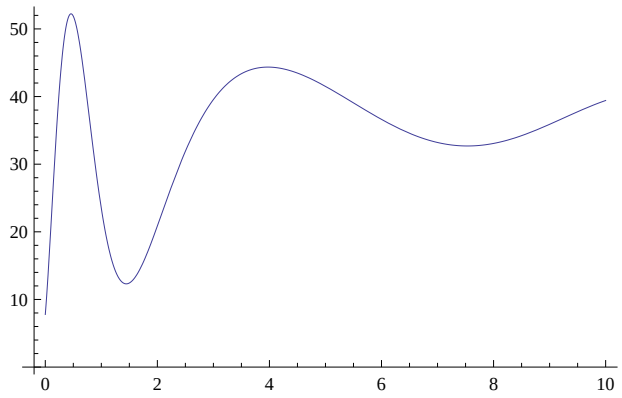
Sumárne zapísané boli dáta pre jednu sériu generované spôsobom:

- $x_1, ..., x_{500}$ náhodne generované z rozdelenia $\mathbf{U}$(0,10)
- $e_1, ..., e_{500}$ náhodne generované z rozdelenia $\mathbf{N}$(0,25)
- $y_1, ..., y_{500}$ $y_i = f_D(x_i) + e_i$, $i$=1,...,500

Pričom funkcia $f_D(x)$ je definovaná:

$$f_D(x) = 10\left(\frac{\sin\left(\frac{x}{10}+0,31\right)^{-2}}{\left(\frac{x}{10}+0,31\right)^{0,8}} + 3 - 0,5\left(\frac{x}{10}+0,31\right)\left(\sin\left(\left(\frac{x}{10}+0,31\right)\pi/4\right)\right)\right), x \in [0, 10]$$

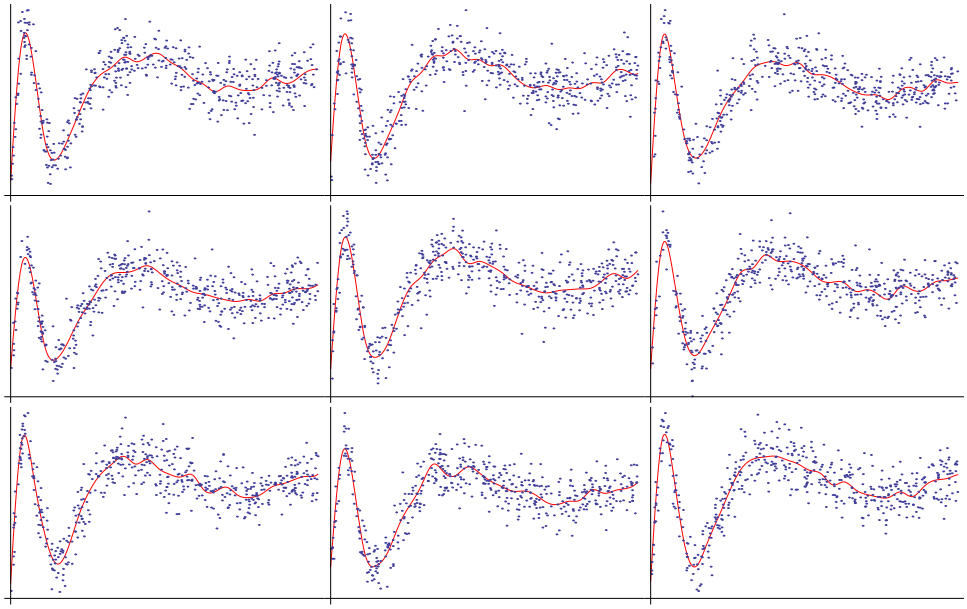
Graf funkcie $f_D$=$f_D(x)$ je ilustrovaný na obrázku 8.




Obr. 8: Graf funkcie $f_D$

Veľké množstvo sérií slúži k tomu, aby boli odhady strednej hodnoty a rozptylu pre dáta ${}_D\lambda_{cv} = ({}_D\lambda_{1,cv}, ..., {}_D\lambda_{u,cv})^{\mathrm{T}}$a ${}_D\sigma^2 = ({}_D\sigma_1^2, ..., {}_D\sigma_u^2)^{\mathrm{T}}$ čo najpresnejšie.

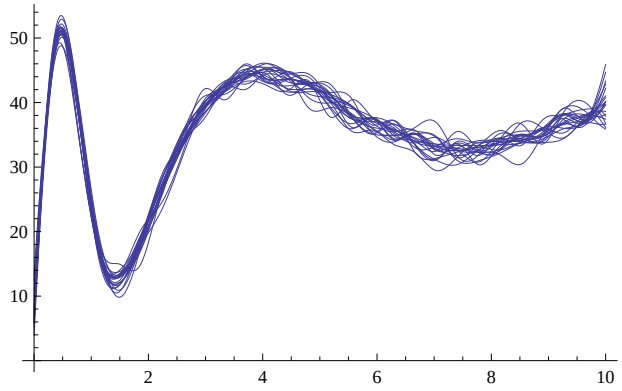
Pre grafickú analýzu metódy je na obrázku 9 uvedených 9 rôznych sérií dát spolu s ich regresnými funkciami.



Obr. 9: P-splajnová regresia pre rôzne série dát

Na jednotlivých grafoch je možné sledovať "správnosť" voľby parametru $\lambda$ metódou GCV. Všetky regresné funkcie, až na malé odchýlky, zodpovedajú tvarom funkcii $f_D$. Ak by bolo $\lambda$ zvolené väčšie, regresná funkcia by bola omnoho hladšia a naopak, v prípade malej hodnoty $\lambda$, by viac zohladňovala jednotlivé náhodné odchýlky. Oba tieto prípady sú nežiadúce a viedli by k tomu, že by regresná funkcia nezodpovedala funkcii $f_D$.

Lepšiu predstavu o rozptyle regresných funkcií poskytne obrázok 10, na ktorom je zo 100 vygenerovaných sérií zobrazených 20 (náhodne zvolených).




Obr. 10: 20 vybraných P-splajnových regresií

V blízkosti krajných bodov $x_l = 0$ a $x_r = 10$ je rozptyl regresných funkcií väčší. Tento jav je spôsobený tým, že v okolí týchto bodov je menej pozorovaní, na základe ktorých je regresná funkcia konštruovaná. Vo všeobecnosti je na grafe "pás" týchto funkcií veľmi tenký.

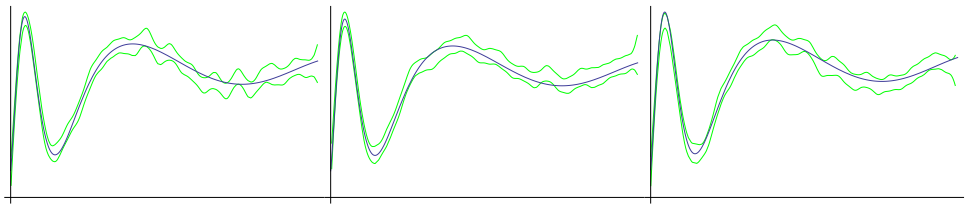
Pre analýzu parametrov boli spočítané výberové stredné hodnoty a rozptyly parametrov celej sady sérií do tabuľky 1.

| | Stredná hodnota | Rozptyl |
|---|---|---|
| ${}_D\lambda_{cv}$ | 0.08053 | 0.00042 |
| ${}_D\sigma^2$ | 24.9677 | 3.04857 |

Tabuľka 1: Výberové stredné hodnoty a rozptyly pre ${}_D\lambda_{cv}$ a ${}_D\sigma^2$

Odhad rozptylu je veľmi presný (skutočný rozptyl $\sigma^2 = 25$) a malá hodnota rozptylu odhadu rozptylu ${}_D\sigma^2$ značí, že takto presný bol pre skoro všetky série dát. Parameter ${}_D\lambda_{cv}$ takmer vôbec nezávisel na konkrétnej sérii pre ktorú bol počítaný.

V poslednej časti simulačnej štúdie bol sledovaný interval spoľahlivosti. Na obrázku 11 sú zobrazené príklady intervalov spoľahlivosti pre tri série a pre $\alpha = 0.05$ spolu s funkciou $f_D$.



$(\alpha = 0.05)$

Obr. 11: Intervaly spoľahlivosti pre 3 série dát

Podľa teórie by pre každý bod $x_i, i = 1, ..., 500$ mal interval spoľahlivosti v tomto bode pokrývať bod $f_D(x_i)$ s pravdepodobnosťou $1-\alpha$. Pre počet dát $N = 500$ v sérii $j$ by teda stredná hodnota počtu bodov $x_i$, v ktorých tento interval nepokrýva hodnotu $f_D(x_i)$ , značíme $M_j$, mala byť: $M_j = 0.05N = 25$. Intervaly spoľahlivosti boli skonštruované pre všetky série dát a bola spočítaná stredná hodnota a rozptyl $M$, kde $M = (M_1, ..., M_{100})^{\mathrm{T}}$.

| | Stredná hodnota | Rozptyl |
|---|---|---|
| $M$ | 25.4432 | 18.9749 |

Tabuľka 2: Stredná hodnota a rozptyl pre veličinu $M$

Grafická analýza a takisto aj analýza odhadov parametrov poskytla údaje, ktoré sa očakávali. Aplikácia metód na simulované dáta tým potvrdila správnosť používaných metód z teórie a ich použitie v praxi bude viesť k požadovaným výsledkom.

# Kapitola 4

# Analýza dát

Táto časť má za cieľ ukázať použitie metód popísaných v minulých kapitolách na reálnych dátach. V úvode budú podrobnejšie popísané samotné dáta, u ktorých sa následne spočítajú základné štatistické ukazovatele. V druhej časti kapitoly sa na dáta aplikuje metóda P-splajnovej regresie a analyzujú sa v krátkosti jej výsledky.

## 4.1 Parvovirus B19

Ako bolo spomenuté v úvode, dáta sa týkajú Parvovirusu B19 (prípadne známeho tiež ako erythrovirus B19). Tento vírus je prvým (a do roku 2005 jediným) vírusom z rodiny parvovírov, ktorým sa môže nakaziť človek. Iné živočíšne druhy vírus nenapadá. Spôsobuje ochorenie *erythema infectiosum*, nazývané aj *piata choroba* (vzhľadom na to, že je piata v poradí chorôb často sa vyskytujúcich u detí). Je to exantematické ochorenie, typicky sa prejavujúce v podobe "zpolíčkovanej tváre".

Piata choroba je len jednou z viacerých prejavov Parvovírusu B19. Nakaziť sa ňou môže človek v každom veku, aj keď najviac je bežná u detí vo veku medzi 6 a 10 rokmi.

Po infikovaní sa choroba u pacienta prejaví po inkubačnej dobe 4 až 14 dní. Prejavuje sa horúčkou a malátnosťou, pokým sa vírus v hojnej miere vyskytuje v krvnom obehu. Pacienti zvyčajne už nie sú infekčnými po tom, čo sa objaví charakteristická vyrážka tejto choroby.

U mladistvých v období puberty a dospelých narozdiel od detí, u ktorých je priebeh choroby mierny, možu nastať bolesti kĺbov rúk, kolien a niekedy aj zápestí. Taktiež, narozdiel od detí, pacienti bývajú stále infekční aj po objavení sa vyrážky.

Parvovírus B19 je často prehliadanou príčinou chronickej anémie u jedincov, ktorí majú ochorenie AIDS. U HIV pozitívnych anemikov bola infekcia parvovírusom B19 v dobe pred zahájením vysoko účinnej antiretrovírovej terapie dokonca najpravdepodobnejšou príčinou anémie. Liečba erythropoetínom a imunoglobulínom podávaným vnútrožilne bola účinná u niektorých pacientov. Infekcia parvovírusom môže spôsobiť silnú reakciu u ľudí s ochorením AIDS, ktorí práve začali s antiretrovírovou terapiou.

Značný nárast výskytu prípadov nastáva každé zhruba 3 až 4 roky, pričom posledná veľká epidémia bola v roku 1998. Ohniská nákazy bývaju predovšetkým v detských jasliach a školách.

Viac o Parvovíruse B19 možno nájsť v Young(2004).

## 4.2 Dáta

Sada dát študovaného Parvovírusu B19 "Parvo B19" obsahuje údaje o pacientoch testovaných na prítomnosť tohoto víru. Zaznamenávaný je vek každého pacienta v čase sérologického testu a hodnota výsledku testu. Tento test meria aktivitu imunoglobulínových

protilátok v krvom obraze ako reakciu pacientovho tela na poslednú infekciu Parvovírusom B19. Výsledné hodnoty zaznamenáva na škále prirodzeného logaritmu. Podrobnejší popis toho, ako tento test funguje, je možné nájsť v práci Hens a kolektív (2008).

Pre účely tejto práce bol použitý výstup z tohoto testu v podobe priemernej hodnoty sérologického testu u $x$-ročného človeka, ktorého vek bol zaokrúhlený nadol. Testu sa zúčastnili pacienti vo veku 0 až 66 rokov.

Základné štatistiky dát sú prezentované v tabuľke 3.

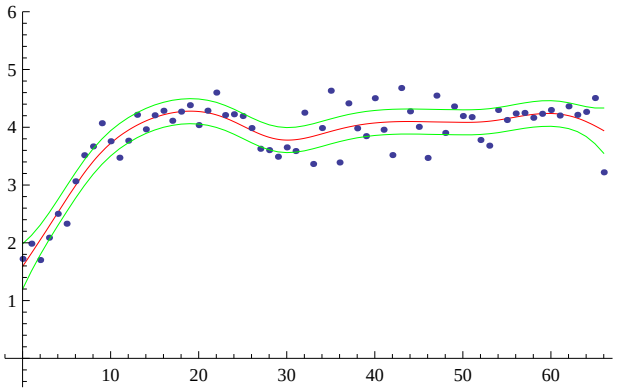
| | Minimum | 1. Kvartil | Medián | Str.hodnota | 3. Kvartil | Maximum |
|---|---|---|---|---|---|---|
| výsledok testu | 1.7017 | 3.6044 | 4.07 | 3.8491 | 4.2662 | 4.6795 |

Tabuľka 3: štatistiky dát "Parvo B19".

Dáta boli získané z Universiteit Hasselt v Belgicku.

## 4.3 Aplikácia P-splajnov

Na modelovanie P-splajnovej regresie sa v tejto sekcii na výpočet $\lambda$ použila metóda generalized cross-validation a na dátach bol skonštruovaný aj interval spoľahlivosti pre $\alpha = 0.05$. Výsledok regresie je na obrázku 12.




$(q = 3, n = 30, k = 2, \alpha{=}0.05)$

Obr. 12: P-splajnová regresia na reálnych dátach

Pre dáta bola spočítaná hodnota $\lambda_{GCV}$ =3.8263.

Z regresnej krivky je možné pozorovať, že u detí s vekom rastie aj hodnota výsledku testu. U dospelých pacientov od veku 20-30 rokov výsledok testu skoro vôbec nezávisí na veku. Interval spoľahlivosti je pri krajných hodnotách $x_l = 0$ a $x_r = 66$ rozšírený. Tento jav je popísaný v kapitole 3.

# Kapitola 5

# Výpočetné prostriedky

Všetky výpočty v tejto práci prebiehali na počítačoch za použitia matematického softvéru. V tejto kapitole bude v krátkosti zhrnuté aké funkcie sa použili a aké problémy môžu nastať pri implementácii metódy P-splajnovej regresie do praxe.

Uvedené metódy (a ostatné pomocné výpočty a grafy) boli implementované v softvéri Mathematica 7. Tento program bol zvoleny kvôli svojej jednoduchosti a zároveň veľkej komplexnosti, kedže ponúka veľkú škálu matematických funkcii. Konkrétne funkcie je možné nájsť v appendixe aj s krátkymi popismi ich fungovania.

Pri použití konkrétnych funkcii z appendixu je potrebné varovať pred ich výpočetnou náročnosťou. Táto náročnosť je dôsledkom hlavne spôsobu výpočtu metódy *cross-validation* a kvôli násobeniu a invertovaniu veľkého počtu mnohorozmerných matíc. Pri zadaní veľkého objemu dát $N$ a veľkého počtu splajnov $n$ rastie táto náročnosť exponenciálne, čo je aj najväčší problém pri implementácii.

Na priloženom CD sú všetky použité funkcie uložené v adresári *Mathematica*.

Zoznam niektorých používaných vstavaných funkcii:

- BSplineBasis - funkcia na konštrukciu B-splajnovej bázy
- LinearSolve - funkcia na riešenie sústavy lineárnych rovníc ( použitá pri výpočtoch P-splajnov)
- Fit - funkcia vhodná na použitie pri metóde najmenších štvorcov ( použitá pri výpočtoch B-splajnov)
- FindMinimum - funkcia na hľadanie minima

# Kapitola 6

# Literatúra


Bollaerts, K., Eilers, P. H. C., and van Mechelen, I. (2006). *Simple and multiple P-splines regression with shape constraints.* British Journal of Mathematical and Statistical Psychology, 59:451-469.

De Boor, C. (1978). *A Practical Guide to Splines.* Springer, Berlin.

Eilers, P. H. C. and Marx, B. D. (1996). *Flexible smoothing with B-splines and penalties.* Statistical Science, 11(2):89-121.

Green, P. J. and Silverman, B. W. (1994). *Nonparameric Regression and Generalized Linear Models.* Chapman and Hall, London.

Hens, N., Aerts, M., Shkedy, Z., Theeten, H., Van Damme, P., and Beutels, P. (2008). *Modelling multi-sera data: the estimation of new joint and conditional epidemiological parameters.* Statistics in Medicine, 27:2651-2664.

Juríček, J., (2009). *Testing for monotonicity of the age-dependent prevalence, based on current status data.*Hasselt University, Hasselt. 13-14.

O'Sullivan, F. (1986). *A statistical perspective on ill-posed inverse problems.* Statistical Science, 1(4):502-527.

Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

Young N S., Brown K E. (2004) . *Parvovirus B19* . National Heart, Lung, and Blood Institute, Bethesda, (350)6:586-97

Zhang, J.-T. (2004). *A simple and efficient monotone smoother using smoothing splines.* Journal of Nonparametric Statistics, 16(5):779-796.

# Dodatok A

# Softvérová implementácia metód

Metódy boli implementované v softvéri *Mathematica*. Ako je spomenuté v 5. kapitole, použité metódy môžu byť výpočetne veľmi náročne. Preto boli všetky procedúry programované so snahou o čo najväčšiu časovú úspornosť, niekedy aj za cenu ich menšej flexibility. V úvode prílohy budú poskytnuté demonštrácie použitia naprogramovaných funkcii. V druhej časti prílohy bude ponúknutý zdrojový kód týchto funkcií, pričom pomocné funkcie nutné k výpočtom budú umiestnené kvôli väčšiemu rozsahu len na CD.

Vo všetkých ponúknutých funkciách je použité rovnaké značenie ako v celej práci a novo definované premenné sú vždy uvedené pri popise danej funkcie. Každá funkcia obsahuje premennú *data*, ktorá značí vstupné dáta $(x_i, y_i)\,, i = 1, ..., N$ , pre ktoré má byť vykonaná daná procedúra.

BSplineApprox[q, n, {xL,xR}, data]

- funkcia spočíta pre dané $q$, $n$, $x_l = xL$ a $x_r = xR$ B-splajnovú regresiu dát

PSplineApprox[q, n, {xL, xR}, diffOperatorDegree, data, lambda]

- funkcia spočíta P-splajnovú regresiu dát pre $k$ =diffOperatorDegree a $\lambda$ =lambda

FindMinCV[q, n, {xL, xR}, diffOperatorDegree, data]

- funkcia určená na výpočet $\lambda$pomocou metódy cross-validation

FindMinGCV[q, n, {xL, xR}, diffOperatorDegree, data]

- funkcia určená na výpočet $\lambda$pomocou metódy generalized cross-validation

PSplineApproxCV[q, n, {xL, xR}, diffOperatorDegree, data]

- funkcia, ktorá spočíta P-splajnovú regresiu dát, pomocou $\lambda_{CV}$

PSplineVariance[q, n, {xL, xR}, diffOperatorDegree, data, lambda]

- funkcia počítajúca odhad rozptylu $\sigma^2$ pre dané $\lambda$

Plot[PSplineApprox[q, n, {xL, xR}, diffOperatorDegree, data, lambda], {x, xL, xR}]

- funkca, ktorá do grafu vykreslí P-splajnovú regresiu dát pre dané $\lambda$

Konkrétna implementácia funkcií do programu Mathematica:

```
BSplineApprox[q_, n_, {xL_, xR_}, data_] := Module[
{basis, f},
basis = GetBSplineBasis[q, n, {xL, xR}];
f = Fit[data, basis, x];
f
];

PSplineApprox[q_, n_, {xL_, xR_}, diffOperatorDegree_, data_, lambda_] :=
Module[
{coeff, splines},
splines = GetBSplineBasis[q, n, {xL, xR}];
coeff =PSplineCoeff[q, n, {xL, xR}, diffOperatorDegree, data, lambda];
coeff.splines
];

PSplineApproxCV[q_, n_, {xL_, xR_}, diffOperatorDegree_, data_] := Module[
{lambda},
lambda = FindMinCV[q, n, {xL, xR}, diffOperatorDegree, data];
PSplineApprox[q, n, {xL, xR}, diffOperatorDegree, data, lambda]
];

PSplineApproxGCV[q_, n_, {xL_, xR_}, diffOperatorDegree_, data_] := Module[
{lambda},
lambda = FindMinGCV[q, n, {xL, xR}, diffOperatorDegree, data];
PSplineApprox[q, n, {xL, xR},
diffOperatorDegree, data, lambda]
];

FindMinCV[q_, n_, {xL_, xR_}, diffOperatorDegree_, data_] := Module[
{f, xmin, points, step, min, max},
points = 16;
step = 0.001;
min = 0.03;
max = min + (points - 1)*step;
f = Interpolation[
Table[{min + step*i,
CV[min + step*i, {q, n, {xL, xR}, diffOperatorDegree,
data}]}, {i, 0, points - 1}]];
xmin = x /. (FindMinimum[f[x], {x, (min + max)/2, min, max}][[2]]);
xmin
];

FindMinGCV[q_, n_, {xL_, xR_}, diffOperatorDegree_, data_] := Module[
{f, xmin, points, step, min, max},
points = 10;
step = 0.01;
min = 0.03;
max = min + (points - 1)*step;
f = Interpolation[
Table[{min + step*i,
GCV[min + step*i, {q, n, {xL, xR}, diffOperatorDegree,
data}]}, {i, 0, points - 1}]];
xmin = x /. (FindMinimum[f[x], {x, (min + max)/2, min, max}][[2]]);
xmin
];
```

```
    PSplineVariance[q_, n_, {xL_, xR_}, diffOperatorDegree_, data_, lambda_] :=
Module[
    {psplain, xData, yData, length = Length[data], A, B, D, H, var},
    psplain =
    PSplineApprox[q, n, {xL, xR}, diffOperatorDegree, data, lambda];
    xData = Table[data[[i, 1]], {i, 1, length}];
    yData = Table[data[[i, 2]], {i, 1, length}];
    (*Print["B"];*)
    B = BMatrix[q, n, {xL, xR}, xData];
    (*Print["D"];*)
    D = DiffOperatorMatrix[n, diffOperatorDegree];
    (*Print["H"];*)
    A = Transpose[B].B + lambda*Transpose[D].D;
    A = Inverse[A];
    H = B.A.Transpose[B];
    var = 1/(length - Tr[H])*Sum[(yData[[i]] - psplain /. x -> xData[[i]])^2, {i, 1, length}];
    var
    ];
```